SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Silvija Vrbančić

LOKALNO PORAVNANJE I PREPOZNAVANJE MOTIVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj, 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem svom mentoru, doc. dr. sc. Pavlu Goldsteinu na ukazanom povjerenju i pruženoj pomoći tijekom izrade diplomskog rada. Također od srca zahvaljujem svojoj obitelji na pruženoj potpori tijekom studija.

Sadržaj

Uvod

Bioinformatika je interdisciplinarna znanost koja primjenjuje matematiku, statistiku i informatiku u biologiji kako bi se analizirale karakteristike živih bića. U području genetike odgovara na pitanja o mutiranju genoma, dok se u strukturalnoj biologiji bavi analizom DNA, RNA i proteina.

U ovom radu analizirat će se proteini i proteomi biljaka kako bi se odgovorilo na pitanje pripadnosti proteina nekoj od proteinske familije. Jedno od važnijih zadaća u bioinformatici je odgovoriti na pitanje o porijeklu proteina, to jest njihovim precima. Kako bismo detaljnije mogli reći da li je neki niz u srodnosti s drugim uvodimo pojam vjerojatnosti. Analiza proteina provodi se uspoređivanjem sličnosti proteina, odnosno lokalnim poravnanjem te su razvijeni mnogi algoritmi koji se bave tim pitanjem. U ovom radu bit će opisan jedan takav algoritam, točnije PSSM algoritam. Njime ćemo dobiti niz "score"-ova koje će biti potrebno analizirati.

Rad je podijeljen na pet poglavlja. U prva dva poglavlja opisujemo matematički aparat koji će nam pomoći u analiziranju podataka. Od čitatelja se pretpostavlja da je upoznat s osnovnim pojmovima u općoj teoriji vjerojatnosti te su stoga u prvom poglavlju navedeni i definirani samo neki osnovni pojmovi koji će se koristiti u radu. Detaljnije analiziranje i dokazi nekih tvrdnji u prvom poglavlju mogu se pronaći u [4]. Nakon kratkog uvoda u teoriju vjerojatnosti uvodimo pojam distribucije ekstremnih vrijednosti čija će se teorija pokazati ključnom u analizi "score"-ova. U tom poglavlju uvedeni su pojmovi poput domene atrakcije i Gumbelove distribucije. Detaljno je obrađena teorija vezana uz Gumbelovu distribuciju dok se više o samoj teoriji ekstremnih vrijednosti može vidjeti u [3]. U trećem poglavlju definira se lokalno poravnanje i "score" poravnanja proteina te je dana definicija proteoma i objašnjene važnosti "score"-ova u proteomu. Nakon toga detaljno je opisan PSSM algoritam. U četvrtom poglavlju dana je statistička analiza "score"-ova na simuliranim podacima. Od čitatelja se također očekuje razumijevanje osnovnih statističkih pojmova poput pripadnosti distribuciji i linearne regresije. Iz tog razloga detalji vezani uz odabir metoda za ispitivanje pripadnosti distribuciji i za odabir regresijskog modela nisu navedeni. Naposljetku, u zadnjem poglavlju su rezultati dobiveni na simuliranim podacima primijenjeni na stvaran proteom biljke *Arabidopsis thaliana*.

Poglavlje 1

Slučajne varijable

Neka je Ω proizvoljan neprazan skup, a $\mathcal{F}$ σ -algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je vjerojatnost na $\mathcal{F}$ ako vrijedi

(i) $\mathbb{P}(A) \geq 0$, $A \in \mathcal{F}$; $\mathbb{P}(\Omega) = 1$.

(ii) $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ za $i \neq j \Rightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definicija 1. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\mathcal{F}$ σ-algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se vjerojatnosni prostor.*

Neka je $\mathscr{B}$ Borelova σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$.

Definicija 2. *Funkcija $X : \Omega \to \mathbb{R}$ je slučajna varijabla ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathscr{B}$*

Neka je $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definirajmo funkciju

$$\mathbb{P}_A(B) = \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad B \in \mathcal{F}.$$

Lako je provjeriti da je $\mathbb{P}_A$ vjerojatnost na $\mathcal{F}$ i nju zovemo uvjetna vjerojatnost uz uvjet A, a $\mathbb{P}(B|A)$ zovemo vjerojatnost od B uz uvjet A.

1.1 Funkcija distribucije. Matematičko očekivanje. Konvergencija.

Skup Ω na kojem je X definirana može biti sasvim općenit, no ako nas zanima problem vezanu za određenu slučajnu varijablu, pogodnije je promatrati vjerojatnosni prostor induciran sa X.

Za $B \in \mathscr{B}$ stavimo

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{\omega \in \Omega; X(\omega) \in B\} = P\{X \in B\} \tag{1.1}$$

Relacijom (1.1) definirana je vjerojatnosna mjera $\mathbb{P}_X : \mathscr{B} \to [0,1]$ koju zovemo vjerojatnosna mjera inducirana sa X, a vjerojatnosni prostor $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ zovemo vjerojatnosni prostor induciran sa X. Prema tome, svakoj slučajnoj varijabli X preko relacije (1.1) na prirodan način se pridružuje vjerojatnosni prostor $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ i problemi vezani za slučajnu varijablu X rješavaju se u okviru tog vjerojatnosnog prostora. Osnovna klasifikacija slučajnih varijabli provodi se na osnovi oblika njihovih funkcija distribucije.

Definicija 3. *Neka je X slučajna varijabla na Ω. Funkcija distribucije od X je funkcija $F_X : \mathbb{R} \to [0,1]$ definirana sa*

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}\left(X^{-1}(-\infty, x]\right) = \mathbb{P}\{\omega \in \Omega; X(\omega) \leq x\} = \mathbb{P}\{X \leq x\}, \quad x \in \mathbb{R}.$$

Često ćemo stavljati $F_X = F$ ako je jasno o kojoj se slučajnoj varijabli radi.

Važno je znati vjerojatnosti događaja vezanih uz X. Način na koji te vjerojatnosti računamo ovisi o tipu slučajne varijable X. Postoje dva glavna tipa slučajnih varijabli: diskretne i neprekidne.

Definicija 4. *Slučajna varijabla X je diskretna ako postoji konačan ili prebrojiv podskup $D \subset \mathbb{R}$ takav da je $\mathbb{P}\{X \in D\} = 1$.*

Definicija 5. *Kažemo da je X apsolutno neprekidna ili, kraće, neprekidna slučajna varijabla ako postoji nenegativna Borelova funkcija f na $\mathbb{R}$ takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t)\, \mathrm{d}\lambda(t), \quad x \in \mathbb{R}. \tag{1.2}$$

Za funkciju distribucije oblika (1.2), odnosno za funkciju distribucije F_X slučajne varijable X kažemo da je apsolutno neprekidna funkcija distribucije. U tom slučaju se funkcija f iz (1.2) zove funkcija gustoće vjerojatnosti od X.

Uvedimo pojam matematičkog očekivanja slučajne varijable X. Neka je X diskretna slučajna varijabla i neka je skup D iz definicije diskretne slučajne varijable, $D = \{x_1, x_2, \ldots\}$ i za svako k vrijedi $\mathbb{P}_X(\{x_k\}) = p_k$. Tada je očekivanje diskretne slučajne varijable X dano sa

$$\mathbb{E}X = \sum_k x_k p_k$$

Neka je sada X neprekidna slučajna varijabla sa funkcijom distribucije F_X, očekivanje slučajne varijable X dano je sljedećom relacijom:

$$\mathbb{E}X = \int_\Omega X \, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} x \, \mathrm{d}F_X(x).$$

Neka je $g : \mathbb{R} \to \mathbb{R}$ Borelova funkcija, vrijedi:

$$\mathbb{E}\left[g(X)\right] = \int_\Omega g(X) \, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g(x) \, \mathrm{d}F_X(x). \tag{1.3}$$

Kako bi uveli i pojam varijance slučajne varijable X, za $r > 0$ definiramo r-ti centralni moment od X:

Definicija 6. *Neka $\mathbb{E}X$ postoji. Tada $\mathbb{E}\left[(X - \mathbb{E}X)^r\right]$ zovemo r-ti centralni moment od X.*

Definicija 7. *Varijanca od X koju označujemo sa VarX ili σ_X^2 je drugi centralni moment od X, dakle*

$$VarX = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Pozitivan drugi korijen iz varijance zovemo standardna derivacija od X i označujemo sa σ_X.

Neka je $(X_n, n \in \mathbb{N})$ niz slučajnih varijabli definiran na istom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$.

Definicija 8. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira gotovo sigurno (g.s) prema slučajnoj varijabli X ako je*

$$\mathbb{P}\left\{\omega \in \Omega : X(\omega) = \lim_{n \to +\infty} X_n(\omega)\right\} = 1.$$

To označujemo sa (g.s.) $\lim_{n\to+\infty} X_n = X$ i takav limes je (g.s.) jedinstven.

Definicija 9. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira po vjerojatnosti prema slučajnoj varijabli X ako za svako $\epsilon > 0$ vrijedi*

$$\lim_{n \to +\infty} \mathbb{P}\left\{|X_n - X| \geq \epsilon\right\} = 0.$$

To označujemo sa $(P)\lim_{n\to+\infty} X_n = X$ i takav limes je također (g.s) jedinstven.

Definicija 10. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira po distribuciji prema slučajnoj varijabli X ako je*

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \quad x \in C(F_X)$$

gdje je $C(F_X)$ skup svih točaka neprekidnosti funkcije F_X.

To označujemo sa $(\mathcal{D})\lim_{n\to+\infty} X_n = X$ ili $X_n \xrightarrow{\mathcal{D}} X$.
Vrijede sljedeće implikacije:

$$(g.s)\lim_{n\to+\infty} X_n = X \Rightarrow (P)\lim_{n\to+\infty} X_n = X \tag{1.4}$$

$$(P)\lim_{n\to+\infty} X_n = X \Rightarrow (\mathcal{D})\lim_{n\to+\infty} X_n = X \tag{1.5}$$

1.2 Primjeri slučajnih varijabli

U ovom dijelu opisat ćemo neke slučajne varijable koje će se pokazati važnima za daljna razmatranja u ovom radu.

Degenerirana slučajna varijabla

Neka je

$$\epsilon(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0, \quad x \in \mathbb{R} \end{cases}$$

Neka je $c \in \mathbb{R}$ proizvoljna konstanta i $F(x) = \epsilon(x - c), x \in \mathbb{R}$. F je očigledno funkcija distribucije. Za slučajnu varijablu X kojoj je F funkcija distribucije kažemo da je degenrirana u točki c. Vrijedi $\mathbb{P}\{X = c\} = 1$.

Uniformno distribuirana slučajna varijabla

Neprekidna slučajna varijabla X ima uniformnu distribuciju na segmentu $[a, b]$, $a, b \in \mathbb{R}$, $a < b$ ako joj je gustoća dana sa

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x < b \\ 0, & x \notin [a, b]. \end{cases} \tag{1.6}$$

Očekivanje i varijanca uniformno distribuirane slučajne varijable X jednake su:

$$\mathbb{E}X = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{a+b}{2},$$

$$\mathrm{Var}X = \frac{(a-b)^2}{12}.$$

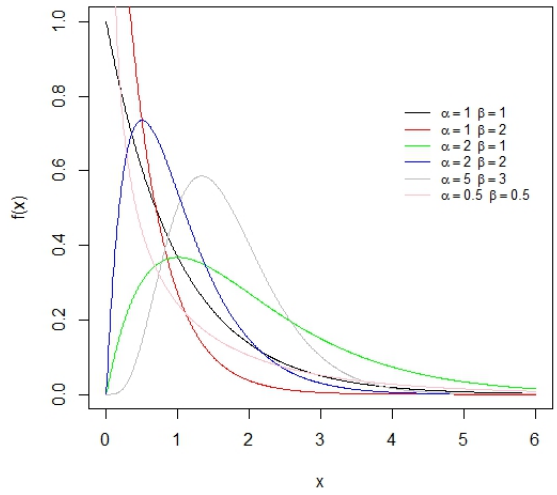
Gama distribucija

Neka je $\alpha > 0, \beta > 0$ i $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$, $x > 0$ gama funkcija. Neprekidna slučajna varijabla X ima gama distribuciju s parametrima α i β ako joj je gustoća f dana sa

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Očekivanje i varijanca gama distribuirane slučajne varijable X jednake su:

$$\mathbb{E}X = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-\frac{x}{\beta}}\, \mathrm{d}x = \alpha\beta, \quad \mathrm{Var}X = \alpha\beta^2. \tag{1.7}$$



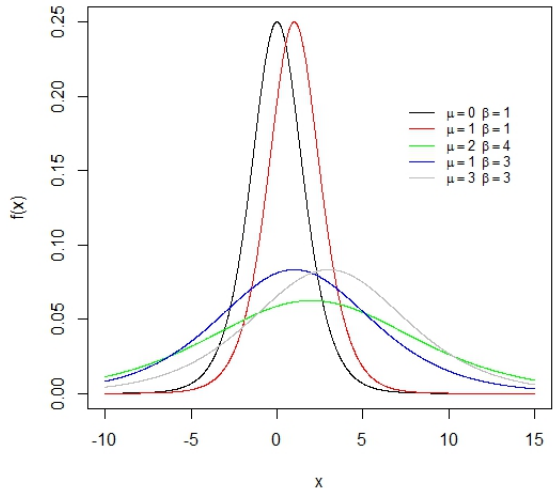
Slika 1.1: Funkcije gustoće gama distribucije s različitim parametrima α i β

Logistička distribucija

Neka je $\mu, \beta \in \mathbb{R}$, $\beta > 0$. Neprekidna slučajna varijabla X ima logističku distribuciju s parametrima μ i β ako joj je funkcija gustoće f dana sa

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta(1 + e^{-\frac{x-\mu}{\beta}})}, \quad x \in \mathbb{R}.$$

Očekivanje slučajne varijable X jednako je $\mathbb{E}X = \mu$, a varijanca $\mathrm{Var}X = \dfrac{\beta^2\pi^2}{3}$.



Slika 1.2: Funkcije gustoće logističke distribucije s različitim parametrima μ i β

Poglavlje 2

Distribucija ekstremnih vrijednosti

Mjerenjem neke pojave od interesa, odnosno prikupljanjem podataka, dobivamo niz vrijednosti koje bismo htjeli analizirati. Osim određivanja funkcije distribucije kojoj bi ta slučajna varijabla mogla pripadati, u primjeni je važnije znati ponašanje njenih ekstremnih vrijednosti. Kako nam za analizu ponašanja ekstremnih vrijednosti nije dovoljno jedno mjerenje, određene zaključke nećemo donositi prema jednoj slučajnoj varijabli već prema nizu $(X_n, n \in \mathbb{N})$ jednako distribuiranih slučajnih varijabli. Svi rezultati će biti iskazani obzirom na maksimum, jer rezultate za minimum dobijemo lako iz relacije:

$$-max(-X) = minX.$$

2.1 Granična distribucija maksimuma i konvergencija prema tipovima

Definicija 11. *Neka su $X_1, X_2, \ldots X_n$ slučajne varijable na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Kažemo da su $X_1, X_2, \ldots X_n$ nezavisne ako za proizvoljne $B_i \in \mathscr{B}$, $(i = 1, 2, \ldots n)$, vrijedi*

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \in B_i\}\right) = \prod_{i=1}^{n}\mathbb{P}\{X_i \in B_i\}.$$

Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih jednako distribuiranih slučajnih varijabli s funkcijom distribucije F. Označimo sa $M_n = \max\{X_1, X_2, \ldots X_n\}$. Ako je $g : \mathbb{R}^n \to \mathbb{R}$ Borelova funkcija, tada će $g(X_1, X_2, \ldots X_n)$ biti slučajna varijabla. Ako definiramo funkciju $g(x) = \max(x), x \in \mathbb{R}^n$, tada vidimo da je tako definirana funkcija g Borelova funkcija. Stoga je

M_n slučajna varijabla. Zanima nas funkcija distribucije od M_n:

$$F_{M_n}(x) = \mathbb{P}\{M_n \leq x\} = \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \leq x\}\right) = \prod_{i=1}^{n}\mathbb{P}\{X_i \leq x\} = \prod_{i=1}^{n}F(x) = [F(x)]^n, \quad x \in \mathbb{R}.$$

Stavimo

$$x_0 = \sup\{x \in \mathbb{R}; F(x) < 1\} \leq \infty. \tag{2.1}$$

Primjetimo da je za $x < x_0$ $F(x) < 1$ pa vrijedi $\mathbb{P}\{M_n \leq x\} = F^n(x) \to 0$. Zbog monotonosti vjerojatnosti slijedi $(P)\lim_{n\to+\infty} M_n = x_0$. Zbog činjenice da je $\{M_n\}$ nerastući niz, konvergencija po vjerojatnosti povlači konvergenciju gotovo sigurno, to jest vrijedi

$$(\text{g.s})\lim_{n\to+\infty} M_n = x_0. \tag{2.2}$$

U primjenama je teško računati sa funkcijom F^n. Zbog toga bismo željeli naći graničnu distribuciju koja će dobro aproksimirati F^n. Iz relacije (2.2) je očito da nedegenerirana granična distribucija neće postojati ukoliko ne normaliziramo M_n. Kao i u većini slučajeva u statistici, normalizaciju najčešće provodimo afinim transformacijama.

Definicija 12. *Kažemo da su dvije funkcije distribucije $V(x)$ i $U(x)$ istog tipa ako za neke $A > 0$ i $B \in \mathbb{R}$ vrijedi $V(x) = U(Ax + B)$, za svaki x.*

Sljedeća propozicija daje eksplicitne formule za granične distribucije koje dobro aproksimiraju F^n. Navodimo ju bez dokaza.

Propozicija 1. *Pretpostavimo da postoje $a_n > 0$, $b_n \in \mathbb{R}$, $n \geq 1$ tako da vrijedi*

$$\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = F^n(a_n x + b_n) \to G(x), \quad \textit{i to slabo kad } n \to \infty,$$

gdje je G nedegenerirana slučajna varijabla. Tada je G funkcija distribucije istog tipa kao jedna od sljedećih funkcija distribucije:

(i) $\Phi_\alpha(x) = \begin{cases} 0, & x < 0 \\ \exp\{-x^{-\alpha}\}, & x \geq 0 \end{cases}$
za neko $\alpha > 0$;

(ii) $\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & x < 0 \\ 1, & x \geq 0 \end{cases}$
za neko $\alpha > 0$;

(iii) $\Lambda(x) = \exp\{-e^{-x}\}, \quad x \in \mathbb{R}$

Φ_α, Ψ_α i Λ zovemo distribucije ekstremnih vrijednosti.

Neka je G distribucija ekstremnih vrijednosti. Kažemo da je F u domeni atrakcije od G (u oznaci $F \in D(G)$) ako postoje $a_n > 0$, $b_n \in \mathbb{R}$, $n \geq 1$ tako da

$$F^n(a_n x + b_n) \to G(x), \quad n \to +\infty, \tag{2.3}$$

gdje je konvergencija u (2.3) slaba konvergencija.
Funkcije Φ_α, Ψ_α i Λ iz Propozicije 1. zvat ćemo redom distribucije Tipa I, Tipa II i Tipa III. Kako će u ovom radu distribucija Tipa III biti od značajnog interesa, slijede neka njezina osnovna obilježja i karakteristike.

2.2 Domena atrakcije distribucije Tipa III

Definicija 13. *Nerastuća funkcija U je Γ-varirajuća (u oznaci $U \in \Gamma$) ako je U definirana na intervalu (x_l, x_0), $x_0 \in \mathbb{R}$, $x_l \in \mathbb{R}$, $x_l < x_0$, $\lim_{x \to x_0} U(x) = \infty$ i ako postoji pozitivna funkcija f definirana na (x_l, x_0) takva da za svaki x vrijedi:*

$$\lim_{t \to x_0} \frac{U(t + xf(t))}{U(t)} = e^x.$$

Funkciju f nazivamo pomoćnom funkcijom.

Neka je $x_0 = \sup\{y : F(y) < 1\}$.

Definicija 14. *Distribucija $F_\#$ čiji je supremum x_0 definiran na gornji način, zovemo von Misesova funkcija ako postoji z_0 takav da je za $z_0 < x < x_0$ i $c > 0$*

$$1 - F_\#(x) = c \exp\Big\{ - \int_{z_0}^{x} (1/f(u))\, \mathrm{d}u \Big\} \tag{2.4}$$

gdje je $f(u) > 0$, $z_0 < u < x_0$ i f je apsolutno neprekidna na (z_0, x_0) sa funkcijom gustoće $f'(u)$ i $\lim_{u \to x_0} f'(u) = 0$. Funkcija f je pomoćna funkcija..

Propozicija 2.

(a) Ako je $F_\#$ von Misesova funkcija definirana kao u (2.4), tada je $F_\# \in D(\Lambda)$. Nizove $(a_n, n \in \mathbb{N})$ i $(b_n, n \in \mathbb{N})$ iz Propozicije 1. možemo definirati na sljedeći način

$$b_n = (1/(1 - F))^{-1}(n)$$
$$a_n = f(b_n),$$

a $1/(1 - F_\#) \in \Gamma$ sa pomoćnom funkcijom f.

(b) Pretpostavimo da je F apsolutno neprekidna i druga derivacija F'' je negativna za sve $x \in (z_0, x_0)$. Ako je

$$\lim_{x \to x_0} F''(x)\,(1 - F(x)) \,/\, (F'(x))^2 = -1, \tag{2.5}$$

tada je F von Misesova funkcija i $F \in D(\Lambda)$. Možemo staviti $f = (1 - F)/F'$. Obratno, von Misesova funkcija koja je dva puta diferencijabilna zadovoljava (2.5).

Za dokaz ove tvrdnje bit će nam potrebna jedna lema i propozicija koje navodimo bez dokaza.

Lema 1. *Pretpostavimo da je $f(u)$ apsolutno neprekidna pomoćna funkcija i $f'(u) \to 0$, $x_0 \to +\infty$. Tada vrijedi*

(a) ako je $x_0 = +\infty$ tada $\lim_{t \to +\infty} t^{-1} f(t) = 0$,

(b) ako je $x_0 < +\infty$ tada $f(x_0) = \lim_{t \to x_0} f(t) = 0$ i $\lim_{t \to x_0} (x_0 - t)^{-1} f(t) = 0$.

U oba slučaja vrijedi

$$\lim_{t \to x_0} (t + x f(t)) = x_0, \quad za\ svex \in \mathbb{R}.$$

Propozicija 3. *Za funkciju distribucije F stavimo*

$$U := 1/(1 - F)$$

tako da je U^{-1} definirana na $(1, \infty)$. Sljedeće dvije tvrdnje su međusobno ekvivalentne:

(a) $F \in D(\Lambda)$

(b) $U \in \Gamma$

Dokaz Propozicije 2.

(a) Neka je $F_\#$ dana sa (2.4), za $x \in \mathbb{R}$ i dovoljno velik t vrijedi:

$$\frac{1 - F_\#(t + x f(t))}{1 - F_\#(t)} = \exp\left\{ - \int_t^{t + x f(t)} (1/f(u))\, \mathrm{d}u \right\} = \exp\left\{ - \int_0^x \{f(t)/f(t + s f(t))\}\, \mathrm{d}s \right\},$$

gdje je $s = (u - t)/f(t)$.
Lema 1. povlači konvergenciju integranta k 1 uniformno na $(0, x)$ te stoga dobivamo

$$\lim_{t \to x_0} \frac{1 - F_\#(t + x f(t))}{1 - F_\#(t)} = e^{-x}.$$

Slijedi da je $1/(1-F_{\#}) \in \Gamma$. Propozicija 3. povlači $F_{\#} \in D(\Lambda)$. Odaberimo $(b_n, n \in \mathbb{N})$ tako da $(b_n, n \in \mathbb{N})$ zadovoljava

$$1 - F_{\#}(b_n) = n^{-1},$$

to jest,

$$b_n = (1/(1 - F_{\#}))^{-1}(n).$$

Zbog toga što je $1/(1 - F(b_n)) \sim n$ slijedi

$$\lim_{n\to\infty} n\,(1 - F_{\#}(b_n + xf(b_n))) = e^{-x}$$

što je ekvivalentno (2.3) nakon logaritmiranja. Stoga vidimo da je za a_n dobar izbor $f(b_n)$.

(b) Stavimo $1 - F = \exp\{-R\}$. Tada će za $f = 1/R'$ i $f' \to 0$ (2.4) vrijediti ako i samo ako $(1/R')' \to 0$. No, $R = -\log(1 - F)$ pa je $R' = F'/(1 - F)$ i $1/R' = (1 - F)/F'$ i

$$(1/R')' = \left(-(F')^2 - (1 - F)F''\right)/(F')^2 = -1 - \left((1 - F)F''\right)/(F')^2$$

što dokazuje tvrdnju. Obrat slijedi trivijalno.

□

Neka je $X \sim \Gamma(\alpha,\beta)$, $F(x)$ njena funkcija distribucije. Tada je $F'(x)$ njena funkcija gustoće, odnosno za $\alpha > 0$ i $\beta > 0$

$$F'(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & x \le 0. \end{cases}$$

Neka je f definirana kao u prethodnoj propoziciji, odnosno $f = (1 - F)/F'$. Uz korištenje L'Hospitalovog pravila imamo

$$\begin{aligned} \lim_{x\to+\infty} f(x) &= \lim_{x\to+\infty} \frac{1 - F(x)}{F'(x)} \\ &= \lim_{x\to+\infty} \frac{-F'(x)}{F''(x)} \\ &= \lim_{x\to+\infty} \frac{-x^{\alpha-1} e^{\frac{-x}{\beta}}}{x^{\alpha-1} e^{\frac{-x}{\beta}} \left((\alpha-1)x^{-1} - \frac{1}{\beta}\right)} \\ &= \lim_{x\to+\infty} \frac{-1}{(\alpha-1)x^{-1} - \frac{1}{\beta}} \\ &= \beta. \end{aligned}$$

Ako stavimo $\beta = 1$ imamo:

$$F''(x) = -F'(x)\left(-(\alpha-1)x^{-1}+1\right) \sim -F'(x) \quad \text{i}$$

$$\lim_{x\to+\infty} F''(x)\,(1-F(x))/(F'(x))^2 = \\ = \lim_{x\to+\infty} -\frac{1-F(x)}{F'(x)} = -1$$

Dakle, prema Propoziciji 2. slijedi $F \in D(\Lambda)$, za $\beta = 1$.
Kako bismo pokazali da se za svaki odabir $\alpha > 0$ i $\beta > 0$ funkcija distribucije gama distribuirane slučajne varijable nalazi u $D(\Lambda)$ moramo naći nizove $(a_n, n \in \mathbb{N})$ i $(b_n, n \in \mathbb{N})$ koji će normalizirati maksimum od n nezavisnih gama distribuiranih slučajnih varijabli. Zbog $\lim_{x\to+\infty} f(x) = 1$, možemo uzeti $a_n = 1$. b_n računamo na sljedeći način:

$$b_n = \left(\frac{1}{1-F}\right)^{-1}(n) \Rightarrow \left(\frac{1}{1-F}\right)(b_n) = n \Rightarrow 1 - F(b_n) = \frac{1}{n},$$

kako vrijedi $1 - F(x) \sim F'(x)$, b_n ćemo dobiti iz relacije $F'(b_n) = \dfrac{1}{n}$. Odnosno,

$$\frac{1}{\Gamma(\alpha)} b_n^{\alpha-1} e^{-b_n} = \frac{1}{n}.$$

Logaritmiranjem dobivamo

$$b_n - (\alpha-1)\log b_n + \log\Gamma(\alpha) = \log n. \tag{2.6}$$

Imajući na umu da $b_n \to +\infty$, dijeljenjem sa b_n dobivamo

$$\lim_{n\to\infty} \frac{\log n}{b_n} = 1,$$

odnosno $b_n \sim \log n$. Ako stavimo $r_n = o(\log n)$, vrijedi $b_n = \log n + r_n$. Uvrstimo li taj izraz za b_n u (2.6), imamo

$$\log n + r_n - (\alpha-1)\log(\log n + r_n) + \log\Gamma(\alpha) = \log n,$$

to jest

$$r_n + \log\Gamma(\alpha) = (\alpha-1)\log\log n + (\alpha-1)\log(1 + r_n/\log n) \\ \Rightarrow r_n = (\alpha-1)\log\log n - \log\Gamma(\alpha) + o(1).$$

Vrijedi

$$b_n - (\log n + (\alpha - 1)\log\log n - \log\Gamma(\alpha))/a_n = o(1)/a_n \to 0$$

te stoga zaključujemo da je

$$b_n = \log n + (\alpha - 1)\log\log n - \log\Gamma(\alpha)$$

prihvatljiv izbor za normalizaciju.

Dakle, za $\alpha, \beta > 0$, $F \in D(\Lambda)$. Odnosno, maksimum n nezavisnih gama distribuiranih slučajnih varijabli ima distribuciju Tipa III.

2.3 Osnovna svojstva funkcije distribucije Tipa III

Funkciju distribucije Tipa III zvat ćemo Gumbelov tip distribucije prema Emilu Gumbelu (1891 – 1966), njemačkom matematičaru koji se bavio modeliranjem ekstremnih događaja u području strojarstva i meteorologije.

Prema (1.2) vrijedi da je funkcija gustoće λ Gumbelove distribucije Λ

$$\lambda(x) = \exp\left\{-x - e^{-x}\right\}, \quad x \in \mathbb{R}.$$

Napišimo Gumbelovu funkciju distribucije u malo općenitijem obliku

$$\Lambda(x) = \exp\left\{-e^{-(x-\mu)/\sigma}\right\}, \ \ \mu, \sigma \in \mathbb{R}, \ \ \sigma > 0. \tag{2.7}$$

Vidimo da uzimanjem $\mu = 0$ i $\sigma = 1$, dobivamo standardnu Gumbelovu funkciju distribucije kao u Propoziciji 1.
Funkcija gustoće tako definirane slučajne varijable glasi

$$\lambda(x) = \sigma^{-1}\exp\left\{-e^{-(x-\mu)/\sigma} - (x-\mu)/\sigma\right\}.$$

Definirajmo slučajnu varijablu $Z := e^{-(X-\mu)/\sigma}$, gdje je X Gumbel distribuirana slučajna varijabla. Z je eksponencijalno distribuirana slučajna varijabla i njena funkcija gustoće je $f_Z(z) = e^{-z}$, $z \geq 0$. Slijedi

$$\mathbb{E}\left[e^{t(X-\mu)/\sigma}\right] = \mathbb{E}\left[Z^{-t}\right] = \Gamma(1-t), \ \ t < 1, \tag{2.8}$$

gdje je Γ gama funkcija definirana kao u 1.2 Primjeri slučajnih varijabli, Poglavlje 1. Iz (4.12) i $t \mapsto t\sigma$ slijedi da

$$\mathbb{E}\left[e^{tX}\right] = e^{t\mu}\Gamma(1-\sigma t), \ \ \sigma|t| < 1.$$

Označimo $\Psi(t) := \log \mathbb{E}\left[e^{tX}\right]$ i $\psi(t) := \dfrac{\Gamma'(t)}{\Gamma(t)}$. Vrijedi

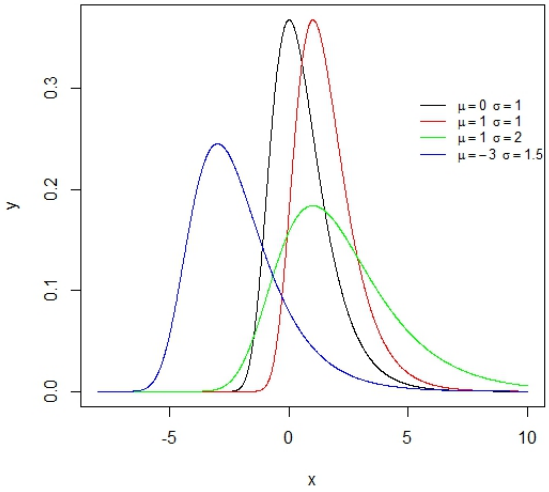
$$\Psi(t) = \mu t + \log \Gamma(1 - \sigma t) \quad \text{i} \quad \Psi'(0) = \mu - \sigma\psi(1).$$

k-ti moment slučajne varijable X jednak je k-toj derivaciji funkcije Ψ u $t = 0$. Stoga vrijedi

$$\mathbb{E}[X] = \mu - \sigma\psi(1) = \mu + \sigma\gamma,$$

gdje je γ Eulerova konstanta ($\gamma \sim 0.5772$), a varijanca je jednaka

$$\mathrm{Var}X = \frac{1}{6}\pi^2\sigma^2.$$



Slika 2.1: Funkcije gustoće Gumbelove distribucije s različitim parametrima μ i σ

Procjena parametara

Znamo da je funkcija gustoće Gumbel distribuirane slučajne varijable s parametrima μ i σ:

$$\lambda(x) = \sigma^{-1} \exp\left\{-e^{-(x-\mu)/\sigma} - (x - \mu)/\sigma\right\}, \quad x \in \mathbb{R}.$$

Neka su $X_1, X_2, \ldots X_n$ nezavisne jednako distribuirane slučajne varijable. Metodom maksimalne vjerodostojnosti želimo procijeniti parametre μ i σ. Odnosno, želimo provesti maksimizaciju log-vjerodostojnosti:

$$l(\mu, \sigma) = -n \log(\sigma) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - \sum_{i=1}^{n} \exp\left\{-\frac{x_i - \mu}{\sigma}\right\}.$$

Maksimum za gornji izraz postiže se za

$$\begin{aligned} \bar{\mu} &= -\sigma \log \frac{1}{n} \sum_{i=1}^{n} \exp\{-n/\sigma\} \quad \text{i} \\ \bar{\sigma} &= \bar{x} - \frac{\sum_{i=1}^{n} (x_i \exp\{-x_i/\sigma\})}{\sum_{i=1}^{n} \exp\{-x_i/\sigma\}}. \end{aligned} \tag{2.9}$$

Poglavlje 3

Lokalno poravnanje

Proteini su makromolekule sastavljene od jednoga ili više lanaca aminokiselina, a unutar živih organizama obavljaju različite funkcije. Proteini su osnovne građevne jedinice živih bića. Služe za ubrzavanje metaboličkih reakcija, transport molekula, umnažanje i prepisivanje DNA, odgovaranje na podražaja i brojne druge funkcije. Proteini su sastavljeni od aminokiselina. Postoji dvadeset standardnih aminokiselina.

Alanin (A)	Arginin (R)	Asparagin (N)	Asparaginska kiselina (D)
Cistein (C)	Glutaminska kiselina (E)	Glutamin (Q)	Glicin (G)
Histidin (H)	Izoleucin (I)	Leucin (L)	Lizin (K)
Metionin (M)	Fenilalanin (F)	Prolin (P)	Serin (S)
Treonin (T)	Triptofan (W)	Tirozin (Y)	Valin (V)

Tablica 3.1: Standardne aminokiseline i njihove kratice

Proteom je skup proteina koje neki organizam proizvodi. Proučavanjem proteoma dobivamo detaljne informacije o živim bićima.

Genetički materijal živih bića se mijenja i tu promjenu nazivamo mutacija. Na proteinskom nivou, mutacije se reflektiraju kao supstitucija (zamjena jedne aminokiseline drugom), insercija (umetanje aminokiseline) i delecija (brisanje aminokiseline). Promotrimo sljedeće mutacije:

```
1  ADNNA
2  DDNNA
3  _DNNA
```

Prijelaz iz niza 1 u niz 2 primjer je za supstituciju, iz 2 u 3 za deleciju, a iz 3 u 2 za inserciju. Kako bi usporedili dva niza aminokiselina, koristimo se poravnanjem. Možemo reći

da je poravnanje rekonstrukcija evolucije. Razlikujemo više vrsta poravnanja. Možemo poravnavati dva niza ili više nizova odjednom. Također, poravnanje može biti globalno i lokalno, ovisno o tome poravnavamo li cijele nizove ili neke njihove podnizove.

3.1 Model ocjenjivanja poravnanja

Kako bismo mogli reći je li neki niz sličan ili različit od drugog niza, potreban nam je model ocjenjivanja poravnanja ("score"). Pretpostavimo da su nam dana dva niza jednakih duljina: `ADNA` i `AANA`. Najjednostavniji primjer ocjenjivanja poravnanja tih nizova bio bi:

(broj mjesta gdje se aminokiseline podudaraju)-(broj mjesta gdje se aminokiseline ne podudaraju).

U gornjem primjeru "score" bi bio $3 - 1 = 2$. Ovako definiran model ocjenjivanja je indikator sličnosti dvaju niza, odnosno uspoređujemo li više proteina s nizom `ADNA`, reći ćemo da je niz najsličniji `ADNA` onaj gdje je "score" maksimalan.

Kako bismo preciznije mogli ustanoviti sličnost dvaju proteina, treba nam malo kompleksniji način definiranja "score"-ova. Neka je R neki slučajni model u kojem nema pretpostavki na ponašanje među aminokiselinama. Neka su $X = x_1x_2x_3 \dots x_{N_1}$ i $Y = y_1y_2y_3 \dots y_{N_2}$ dva proteina. Neka je q_{x_i} vjerojatnost da se x_i pojavi u nizu X, a q_{y_i} vjerojatnost da se y_i pojavi u nizu Y. Pretpostavljamo da je pojava svake aminokiseline u nizu nezavisna, odnosno pojava jedne aminokiseline ne utječe na vjerojatnost pojavljivanja neke druge aminokiseline. U tom slučaju će vjerojatnost da se nizovi X i Y pojave biti jednaka

$$\mathbb{P}(X, Y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}. \tag{3.1}$$

Neka je sada M model u kojem se parovi aminokiselina pojavljuju s vjerojatnošću p_{x_i,y_i}. Vjerojatnost p_{x_i,y_i} može biti shvaćena kao vjerojatnost da su se aminokiseline x_i i y_i nezavisno pojavile iz neke treće nepoznate aminokiseline z u procesu mutacije, to jest z je jednaka x_i i/ili y_i. Tada je vjerojatnost poravnanja nizova X i Y

$$\mathbb{P}(X, Y \mid M) = \prod_i p_{x_i,y_i}. \tag{3.2}$$

Omjer vjerojatnosti u (3.2) i (3.1) naziva se omjer šansi (eng. odds ratio) i jednak je:

$$\frac{\mathbb{P}(X, Y \mid M)}{\mathbb{P}(X, Y \mid R)} = \frac{\prod_i p_{x_i,y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}.$$

Kako bismo lakše mogli analizirati prethodni omjer, htjeli bismo aditivni sistem ocjenjivanja pa ćemo gornji omjer logaritmirati i označiti sa S, odnosno

$$S = \sum_i s(x_i, y_i) \quad \text{gdje je} \quad s(x_i, y_i) = \log \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}. \tag{3.3}$$

$s(x_i, y_i)$ je logaritam omjera vjerojatnosti da se aminokiseline x_i i y_i pojave kao poravnat par i vjerojatnosti da x_i i y_i nisu poravnate.

Pretpostavimo da su duljine nizova X i Y različite. Označimo sa N_1 duljinu niza X, a sa N_2 duljinu niza Y. Pretpostavimo da je $N_1 < N_2$. Neka je $X = x_1x_2x_3 \dots x_{N_1}$ a $Y = y_1y_2y_3 \dots y_{N_2}$. Izračunamo "score" poravnanja između podniza niza Y i X. Shemu računanja "score"-ova možemo ilustrativno prikazati na sljedeći način.

$$\begin{array}{ll} Y & y_1\ y_2\ y_3\ \dots\ y_{N_1}\ \dots\ y_{N_2} \\ X & x_1\ x_2\ x_3\ \dots\ x_{N_1} \end{array}$$

$$\begin{array}{ll} Y & y_1\ y_2\ y_3\ \dots\ \dots\ y_{N_1+1}\ \dots\ y_{N_2} \\ X & \quad x_1\ x_2\ x_3\ \dots\ x_{N_1} \end{array}$$

$$\vdots$$

$$\begin{array}{ll} Y & y_1\ \dots\ \dots\ \dots\ y_{N_2-N_1}\ \dots\ \dots\ y_{N_2-1}\ y_{N_2} \\ X & \quad\quad\quad\quad\quad\quad x_1\ x_2\ x_3\ \dots\ x_{N_1} \end{array}$$

$$\begin{array}{ll} Y & y_1\ \dots\ \dots\ \dots\ y_{N_2-N_1}\ y_{N_2-N_1+1}\ \dots\ \dots\ y_{N_2-1}\ y_{N_2} \\ X & \quad\quad\quad\quad\quad\quad\quad\quad x_1\ x_2\ x_3\ \dots\ x_{N_1-1}\ x_{N_1} \end{array}$$

Preciznije, za svaku odabranu poziciju $i \in \{1, 2 \dots, N_2 - N_1 + 1\}$ u nizu Y, poravnamo niz $y_iy_{i+1} \dots y_{N_1+i-1}$ sa nizom X. Sveukupno ćemo imati $N_2 - N_1 + 1$ mogućih poravnanja i $N_1 (N_2 - N_1 + 1)$ usporedba aminokiselina. Ovaj način računanja "score"-ova naziva se klizeći prozor. Nakon što izračunamo $N_2 - N_1 + 1$ "score"-ova, uzimamo onaj maksimalan kao indikator vjerojatnosti da se niz X nalazi u nizu Y.

Strukturalni motivi su kratki segmenti proteinskih struktura te imaju strukturalnu ili funkcionalnu ulogu. Prisutnost nekog motiva u proteinu ukazuje na njegovu ulogu te se na taj način proteini klasificiraju u proteinske familije. Kod računanja "score"-ova taj motiv nazivat ćemo upitom. Vjerojatnost da se neki motiv pojavljuje u proteinu računati

ćemo pomoću "score"-ova na prethodno opisan način uspoređivanja dvaju nizova proteina različite duljine. Zanimat će nas distribucija tih "score"-ova te distribucija maksimalnih "score"-ova u cijelom proteomu.

3.2 Simulacija proteoma

Neka je $\mathfrak{A} = (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V)$. Uzmimo da je $P = (p_i)$, $i = 1, 2, \ldots, 20$ vjerojatnost pojavljivanja i-te po redu aminokiseline iz $\mathfrak{A}$, tako da je $\sum_{i=1}^{20} p_i = 1$. Neka je $s(1) = p(1)$, a $s(i) = s(i-1) + p(i)$ za $i = 2, \ldots, 20$. Neka je a slučajan realan broj iz $[0, 1)$ i $i \in \{1, 2, \ldots 20\}$ najmanji indeks za kojeg vrijedi $a \leq s(i)$. Pretpostavimo da želimo simulirati protein koji sadrži n aminokiselina. Tada ćemo n puta odabrati slučajan broj iz $[0, 1)$ i odabrati i-tu aminokiselinu te je dodati u niz. Proteomi biljaka imaju nekoliko tisuća nizova prosječne duljine nekoliko stotina. Analize koje će se vršiti u daljnem radu bit će temeljene na dvije simulacije proteoma:

1) proteom ima 30000 proteina, a svaki niz sadrži točno 1000 aminokiselina,

2) proteom ima 30000 proteina, a duljine proteina dolaze iz uniformne distribucije $U(50, 3000)$.

Poglavlje 4

PSSM

Neka je R slučajan model i $Q = (q_i)$, $i = 1, \ldots, 20$, stohastički vektor definiran kao u (3.1), a parametri modela M definiranog u (3.2) neka su dani u matrici $P = (M_{i,j})$, $i = 1, \ldots N_1$, $j = 1, \ldots, 20$ gdje je N_1 duljina upita. Kako bismo odredili model M, želimo naći varijante nekog enzima u nekom novom organizmu. Promotrimo nekoliko varijanti, odnosno varijacija nekog motiva:

```
A V G S D
F V G S N
A L G S D
F L G S D
```

Računamo relativne frekvencije pojavljivanja aminokiselina u gornjim varijacijama. Vidimo da se na prvim mjestima aminokiseline `A` i `F` pojavljuju jednak broj puta. Stoga će element $M_{1,1}$ i $M_{1,14}$ u matrici P biti jednaki $\frac{1}{2}$, a ostali elementi bit će jednaki nuli. Na analogan način izračunamo preostale elemente matrice P. Iz tog razloga ovaj algoritam nazivamo PSSM (eng. position specific scoring matrix).

Neka je

$$\begin{aligned} p_k &= \mathbb{P}(y_{k+1} \mid M_1)\, \mathbb{P}(y_{k+2} \mid M_2) \cdots \mathbb{P}(y_{k+N_1} \mid M_{N_1}) \\ &= \prod_{j=1}^{N_1} \mathbb{P}\left(y_{k+j} \mid M_j\right), \quad k = 0, 1, \ldots, N_2 - N_1. \end{aligned}$$

Za $k \in \{0, 1, \ldots, N_2 - N_1\}$, p_k je vjerojatnost poravnanja niza $y_k, y_{k+1}, \ldots y_{k+N_1}$ sa nizom koji je opisan modelom M. Tada je vjerojatnost da se niz Y poravna prema modelu M jednaka

$$\mathbb{P}(Y \mid M) = \max_{k \in \{0,1,\ldots,N_2-N_1\}} p_k.$$

Sa $q_{y_{k+j}}$ označimo element vektora Q na mjestu i na takav način da je i indeks aminokiseline y_{k+j} u $\mathfrak{A}$. Omjer šansi da se niz $y_k, y_{k+1}, \ldots y_{k+N_1}$ poravna s nizom koji je opisan s modelom M dan je sa

$$\prod_{j=1}^{N_1} \frac{\mathbb{P}\left(y_{k+j} \mid M_j\right)}{q_{y_{k+j}}}, \quad k \in \{0, 1, \ldots, N_2 - N_1\}.$$

Tada je prema (3.3), "score" poravnanja S_k niza $y_k, y_{k+1}, \ldots y_{k+N_1}$ i niza opisanog sa M

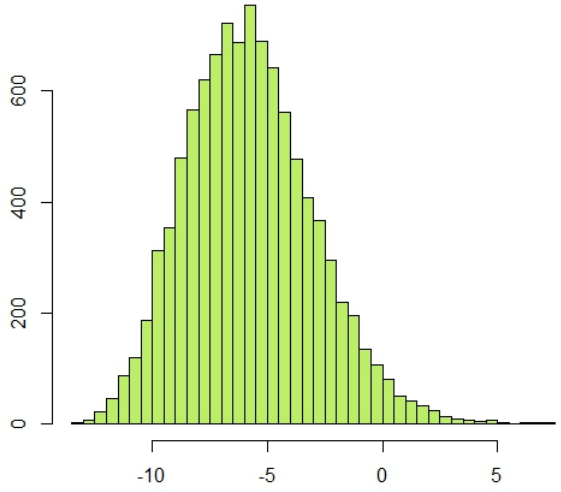
$$S_k = \sum_{i=1}^{N_1} \log \frac{\mathbb{P}\left(y_{k+j} \mid M_j\right)}{q_{y_{k+j}}}, \quad k \in \{0, 1, \ldots, N_2 - N_1\}. \tag{4.1}$$

Omjer šansi S da se niz Y poravna prema M računat ćemo kao maksimum "score"-ova S_k, odnosno

$$S = \max_{k \in \{0,1,\ldots,N_2-N_1\}} S_k. \tag{4.2}$$

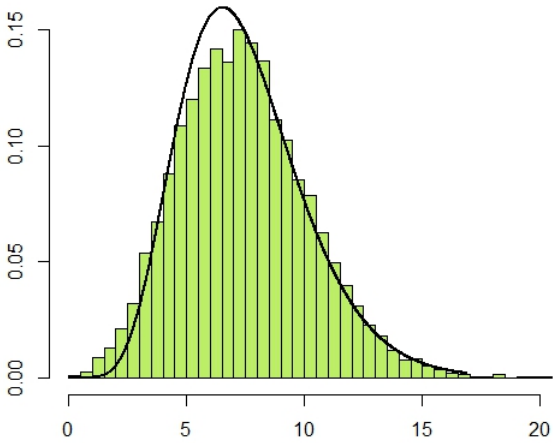
4.1 Distribucija "score"-ova

U simuliranom proteinu duljine 1000 želimo analizirati sve "score"-ove koje dobivamo poravnanjem s `ATWYVRILKLNATWYV`. Kako je duljina upita 16, vidimo da ćemo sveukupno dobiti $1000 - 16 + 1 = 985$ "score"-ova izračunatih kako je opisano sa (4.1). Distribucija tih "score"-ova prikazana je na sljedećoj slici:



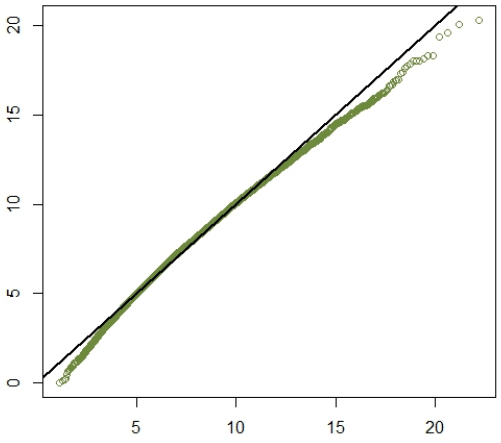
Slika 4.1: Histogram "score"-ova

Možemo naslutiti da ovakvi podaci slijede gama razdiobu. No kako je gama distribuirana slučajna varijabla definirana na $(0, +\infty)$, "score"-ovima pribrojimo apsolutnu vrijednost minimuma. Neka je $\bar{X}$ sredina tih "score"-ova, a S^2 uzoračka varijanca. Iz (1.7) slijedi $\bar{\beta} = S^2/\bar{X}$, a $\bar{\alpha} = \bar{X}/\bar{\beta}$. Dobivamo $\bar{\alpha} = 0.1380328$ i $\bar{\beta} = 1.022282$. Na sljedećoj slici prikazani su podaci i funkcija gustoće $\Gamma(\bar{\alpha}, \bar{\beta})$.



Slika 4.2: Histogram i funkcija gustoće $\Gamma(\bar{\alpha},\bar{\beta})$

Usporedimo li kvantile gama distribucije i dane podatke dobivamo sljedeći $Q-Q$ graf gdje su na x-osi prikazani kvantili podataka, a na y-osi kvantili teoretske, odnosno gama distribucije.



Slika 4.3: $Q-Q$ graf

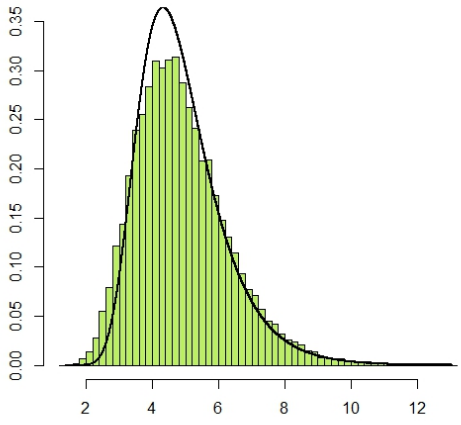
Vidimo da se se kvantili dobro grupiraju oko pravca $y = x$, pa možemo reći da "score"-ovi slijede gama distribuciju.

4.2 Distribucija maksimalnih "score"-ova

Proteom jednakih duljina nizova

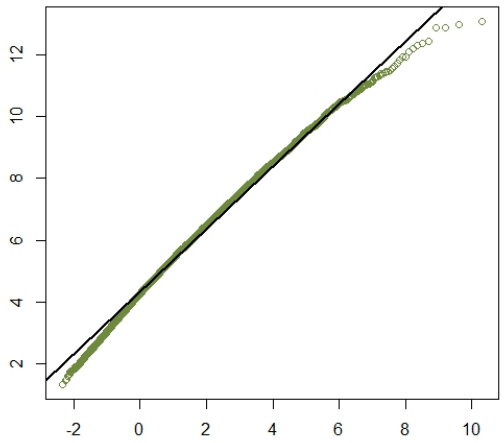
Simuliramo proteom koji sadrži 30000 nizova duljine 1000. Simulirani proteini su međusobno nezavisni i "score"-ovi u nizovima dobiveni PSSM-om su gama distribuirani. Kako bismo mogli reći nalazi li se motiv `ATWYVRILKLNATWYV` u nekom od 985 podnizova u simuliranom proteinu, pogledamo najveći mogući "score" u tom nizu na način kako je opisano sa (4.2). Naime, od interesa je znati da li se neki određeni enzim, odnosno motiv, nalazi u proteomu. Stoga nas zanimaju samo "score"-ovi dobiveni kao maksimalni "score"-ovi po nizovima koji čine proteom. U ovom simuliranom proteomu dobit ćemo 30000 "score"-ova izračunatih prema (4.2). Kako "score"-ovi u nizovima dolaze iz gama distribucije, a gama distribuirana slučajna varijabla je u domeni atrakcije Gumbel distribuirane slučajne varijable, 30000 maksimalnih "score"-ova trebalo bi slijediti Gumbelovu distribuciju.

Procijenimo parametre μ i σ Gumbelove distribucije iz maksimalnih "score"-ova. Iz (2.9) slijedi da je $\bar{\mu} = 4.324313$ i $\bar{\sigma} = 1.00964$. Na sljedećoj slici prikazan je histogram maksimalnih "score"-ova i funkcija gustoće Gumbel distribuirane slučajne varijable s procijenjenim parametrima $\bar{\mu}$ i $\bar{\sigma}$:



Slika 4.4: Histogram i funkcija gustoće Gumbelove slučajne varijable

Usporedimo li kvantile Gumbelove distribucije i dane podatke dobivamo sljedeći $Q-Q$ graf gdje su na x-osi prikazani kvantili podataka, a na y-osi kvantili teoretske distribucije.

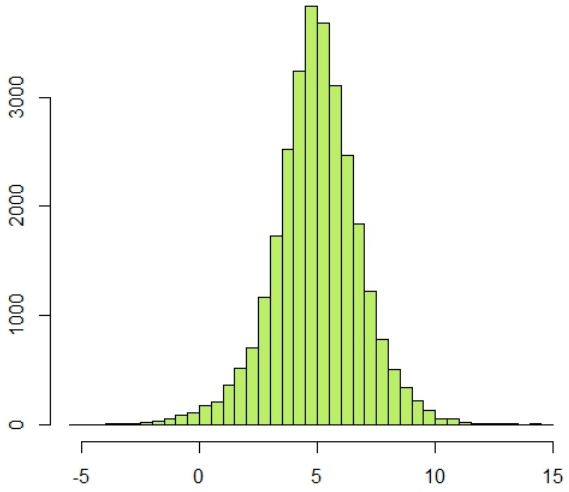


Slika 4.5: $Q-Q$ graf

Vidimo da se se kvantili dobro grupiraju oko pravca $y = \bar{\mu} + \bar{\sigma}x$, pa možemo reći da maksimalni "score"-ovi slijede Gumbelovu distribuciju, kao što se očekivalo.

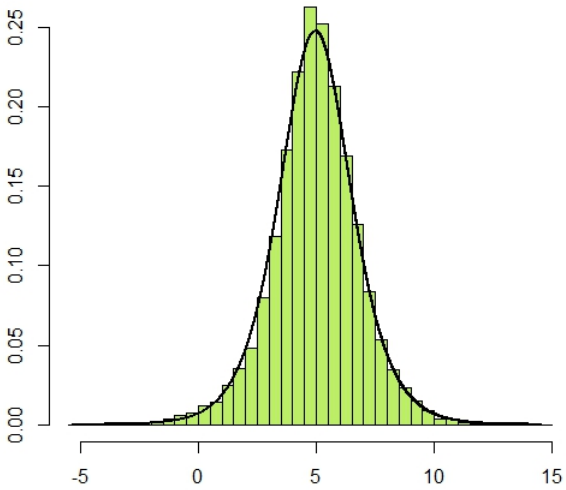
Proteom nejednakih duljina nizova

Simuliramo proteom čiji proteini imaju duljine koje slijede uniformnu distribuciju $U(50, 3000)$. Kao i u prošlom poglavlju, zanima nas distribucija maksimalnih "score"-ova. Na sljedećoj slici prikazan je histogram maksimalnih "score"-ova dobivenih PSSM algoritmom sa istim upitom `ATWYVRILKLNATWYV`.



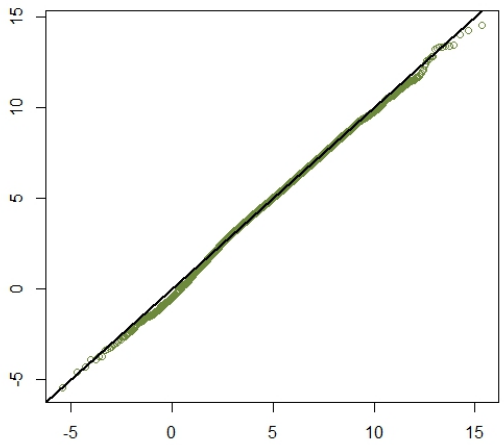
Slika 4.6: Histogram "score"-ova

Već se iz histograma može naslutiti da maksimalni "score"-ovi u proteomu nejednakih duljina nizova neće biti Gumbel distribuirani. Pokaže se da maksimalni "score"-ovi u ovom slučaju imaju logističku razdiobu. Prema (1.2) dobivamo procijenjene parametre logističke distribucije $\bar{\alpha} = \mathbb{X}$ i $\bar{\beta} = \dfrac{\sigma\sqrt{3}}{\pi}$, odnosno $\bar{\alpha} = 4.97146$ i $\bar{\beta} = 1.009597$.



Slika 4.7: Histogram i funkcija gustoće logističke razdiobe

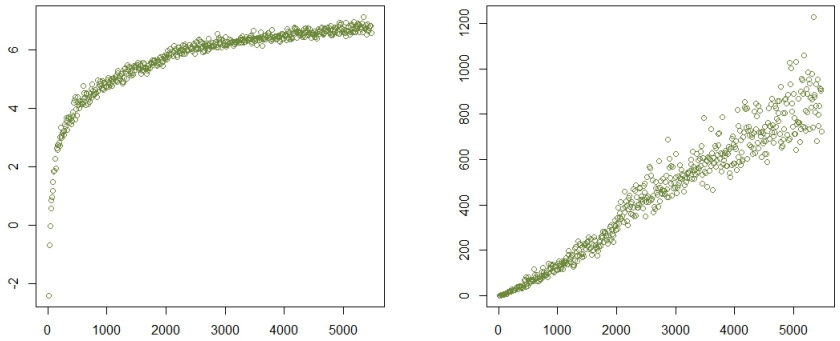
Na slici 4.8 prikazan je histogram maksimalnih "score"-ova i funkcija gustoće logističke razdiobe s parametrima $\bar{\alpha}$ i $\bar{\beta}$. Iz sljedećeg $Q-Q$ grafa možemo vidjeti da ti "score"-ovi zaista slijede logističku razdiobu.



Slika 4.8: $Q-Q$ graf

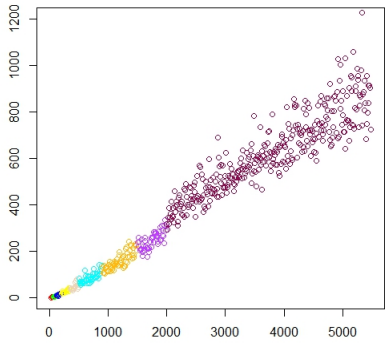
4.3 Korekcije na duljinu

Kako su "score"-ovi u proteomu jednakih duljina nizova Gumbel distribuirani, a u proteomu s nejednakim duljinama nizova slijede logističku distribuciju, može se zaključiti da "score"-ovi ovise o duljini niza. Kako bismo tu ovisnost lakše uočili, simuliramo proteom koji ima 54700 proteina tako da su prvih 100 nizova duljine 20, drugih 100 nizova duljine 30, itd. Za dani upit izračunamo PSSM "score"-ove za istu duljinu niza. Kako rezultati ne bi ovisili o jednom proteinu, izračunamo aritmetičku sredinu "score"-ova po duljinama. Na taj način dobivamo 547 "score"-ova koji redom pripadaju duljinama $20, 30, 40, \ldots, 5480$.



Slika 4.9: "score"-ovi u ovisnosti o duljinama

Na slici 4.9 (lijevo) prikazani su prosječni "score"-ovi u ovisnosti o duljinama. Kako smo "score"-ove računali kao logaritme, za bolji uvid ih eksponenciramo.Tada dobivamo podatke prikazane na slici 4.9 (desno). Već se iz grafičkog prikaza može vidjeti da će se kod procjene eksponenciranih prosjeka metodom najmanjih kvadrata pojaviti heteroskedastičnost, odnosno promjena varijance s rastom duljine niza. Iz tog razloga prosjeke, odnosno podatke podijelimo na dijelove obzirom na duljinu tako da varijabilnost po dijelovima bude manja. Jedan način za podjelu je sljedeći:



Slika 4.10: Eksponencirani "score"-ovi

Takvom podjelom dobivamo 9 skupina podataka. Za svaku skupinu metodom najmanjih kvadrata procijenimo polinom koji najbolje opisuje podatke u ovisnosti o duljini niza iz koje prosjeci dolaze. Označimo sa n duljinu niza. Rezultati su sljedeći:

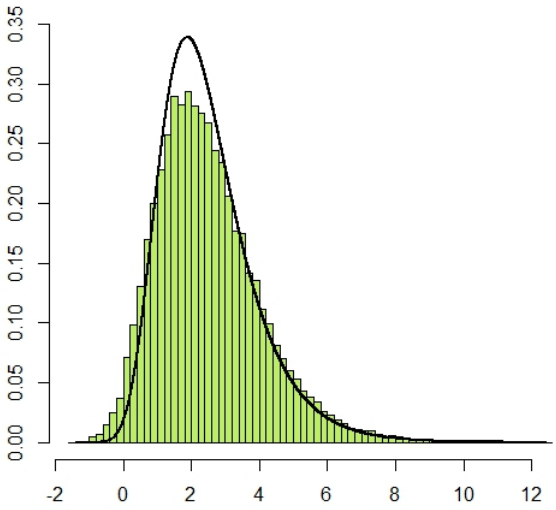
Duljine	Polinom
< 60	$-1.161 \cdot 10^{-4}n + 8.211 \cdot 10^{-6}n^2$
$[60, 120)$	$8.865 \cdot 10^{-6}n^2$
$[120, 200)$	$1.068 \cdot 10^{-5}n^2$
$[200, 350)$	$1.138 \cdot 10^{-5}n^2$
$[350, 500)$	$1.197 \cdot 10^{-5}n^2$
$[500, 900)$	$9.612 \cdot 10^{-6}n^2$
$[900, 1500)$	$9.455 \cdot 10^{-6}n^2$
$[1500, 2000)$	$1.927 \cdot 10^{-1}n - 2.616 \cdot 10^{-5}n^2$
≥ 2000	$3.268 \cdot 10^{-6}n - 1.007 \cdot 10^{-9}n^2 + 9.893 \cdot 10^{-14}n^3$

Tablica 4.1: Parametri regresija

Kako su ove korekcije izračunate u ovisnosti o eksponencijalnim prosjecima, "score"-ovima izračunatim PSSM-om moramo oduzeti logaritam vrijednosti korekcija za dane duljine.

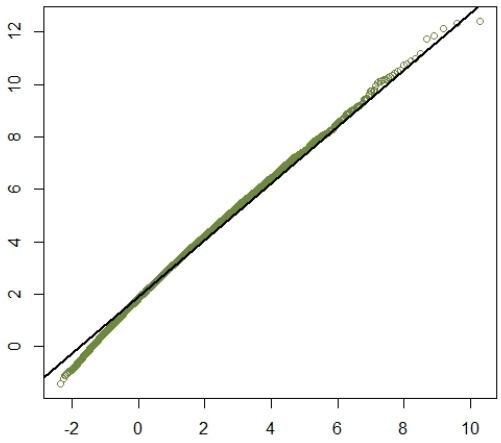
Distribucija "score"-ova uz korekcije

Promotrimo opet simulirani proteom nejednakih duljina proteina. Maksimalni "score"-ovi u tom proteomu slijede logističku distribuciju (Slika 4.8). Nakon korekcija očekujemo da "score"-ovi slijede Gumbelovu distribuciju. Procijenimo li parametre Gumbelove distribucije dobivamo $\bar{\mu} = 1.885298$ i $\bar{\sigma} = 1.084358$. Histogram "score"-ova i funkcija gustoće Gumbelove distribucije s parametrima $\bar{\mu}$ i $\bar{\sigma}$ prikazana je na sljedećoj slici.



Slika 4.11: Histogram i funkcija gustoće Gumbelove razdiobe

Uspoređivanjem kvantila Gumbelove distribucije i podataka dobiva se $Q-Q$ graf prikazan na slici 4.12. Vidimo da "score"-ovi zaista slijede Gumbelovu razdiobu. Dakle, uvođenjem korekcija na "score"-ove koji su logistički distribuirani, dobivamo Gumbel distribuirane "score"-ove.



Slika 4.12: $Q - Q$ graf

Poglavlje 5

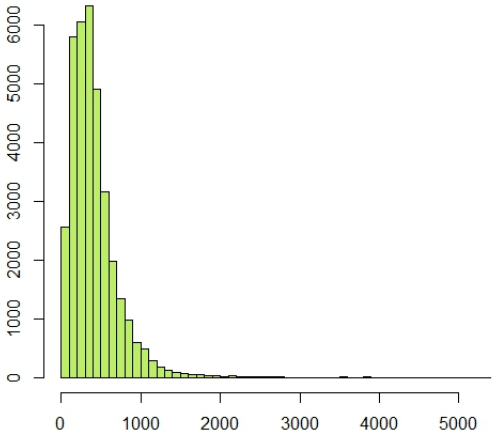
Proteom biljke *Arabidopsis thaliana*

Arabidopsis thaliana je mala biljka s cvijetovima koja često služi kao model za istraživanje u biologiji. Pripada familiji *Brassicaceae*, u koju pripadaju i neke kultivirane vrste poput kupusa i rotkvice.

Slika 5.1: *Arabidopsis thaliana*

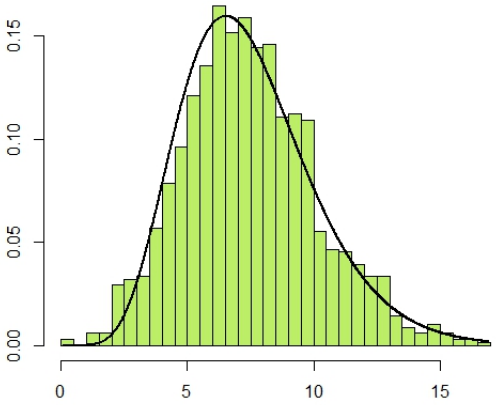
Iako nije od značajne važnosti za agronomiju, *Arabidopsis* je pogodna za istraživanja u genetici i molekularnoj biologiji. Naime, postupkom sekvenciranja aproksimativno je dobiveno 115Mb proteoma od 125Mb, genetička mapa svih 5 kromosoma je poznata i ima kratak životni ciklus od približno 6 tjedana.

Aproksimativni proteom *Arabidopsis thaliana*-e na kojem ćemo računati PSSM "score" -ove ima 35176 nizova čije su duljine prikazane na sljedećoj slici.

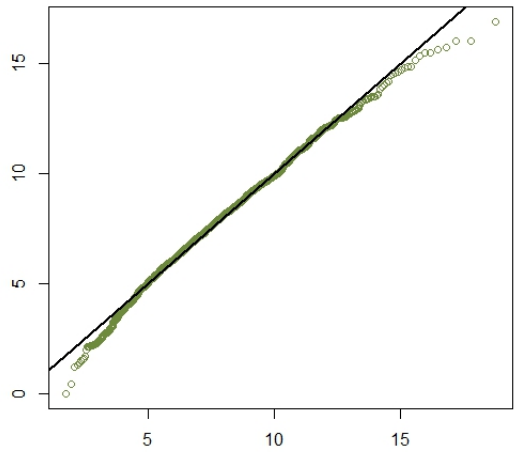


Slika 5.2: Histogram duljina

Distribucija "score"-ova dobivenih PSSM-om kao što je opisano sa (4.1) s upitom `ATWYVRILKLNATWYV` u jednom proteinu također slijedi gama distribuciju. Na sljedećim slikama prikazan je histogram "score"-ova s funkcijom gustoće gama distribucije s procijenjenim parametrima i $Q - Q$ graf kvantila gama distribucije i podataka.

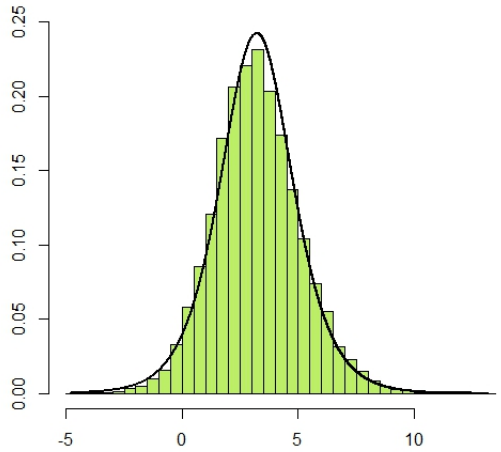


Slika 5.3: Histogram i gama razdioba

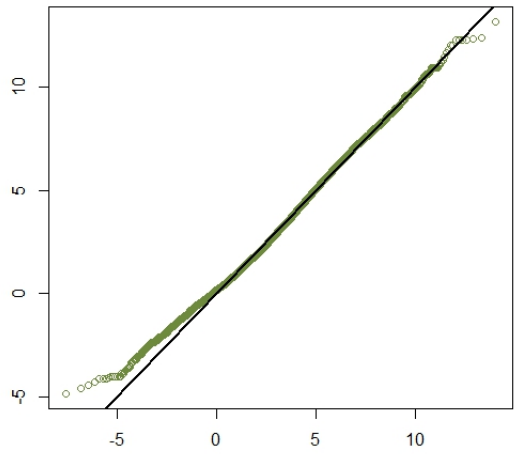


Slika 5.4: $Q - Q$ graf

Distribucija maksimalnih "score"-ova slijedi logističku distribuciju kao što je bio slučaj i u simuliranom proteomu. (Slika 5.5 i Slika 5.6)

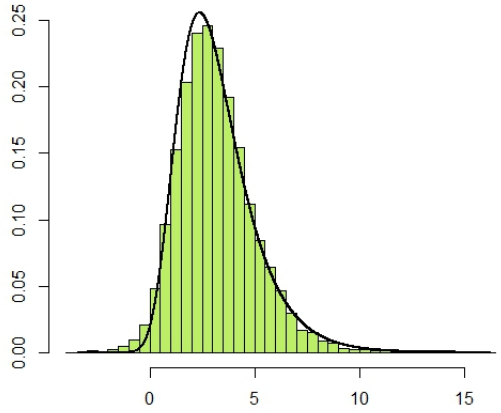


Slika 5.5: Histogram i logistička razdioba



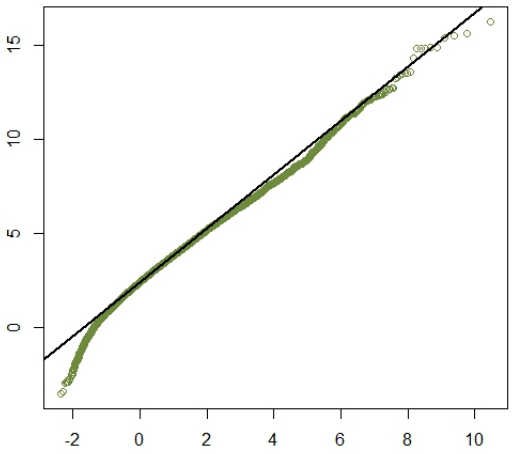
Slika 5.6: $Q - Q$ graf

Na slici 5.7 prikazani su "score"-ovi nakon korekcija danih u Tablici 4.1 i Gumbelova funkcija gustoće dobivena procjenom parametra iz korigiranih "score"-ova.



Slika 5.7: Histogram i Gumbelova razdioba

Uspoređivanjem kvantila na slici 5.8 vidimo da maksimalni "score"-ovi u proteomu biljke *Arabidopsis thaliana*-e nakon korekcija zaista dolaze iz Gumbelove distribucije.



Slika 5.8: $Q - Q$ graf

Bibliografija

[1] Richard Durbin, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press, 1998.

[2] Warren John Ewens i Gregory R Grant, *Statistical methods in bioinformatics: an introduction*, sv. 746867830, Springer, 2005.

[3] Sidney I Resnick, *Extreme values, regular variation, and point processes*, Springer, 2007.

[4] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

Sažetak

Primjenjujući PSSM algoritam na jedan protein dobivamo “score”-ove koji su gama distribuirani. Pokazano je da maksimumi nezavisnih gama distribuiranih slučajnih varijabli slijede Gumbelovu distribuciju. U slučaju kad su nizovi u proteomu jednake duljine, dobivamo Gumbelovu distribuciju, no kada su nejednake duljine dobivamo logističku. Uklanjanjem ovisnosti “score”-ova o duljini, logistički distribuirani “score”-ovi prelaze u Gumbel distribuirane.

Summary

In this work, we show that sliding window algorithm yields gamma distributed scores (for a protein of sufficient length). We work through the theoretical background to show that the maxima of independent and identically distributed gamma random variables are Gumbel distributed. We then analyze the distribution of scores for sequences of unequal length, show that it follows logistic distribution, and construct the length-correction, that should yield Gumbel distributed scores.

Životopis

Rođena sam 26.06.1990. u Zagrebu. Svoje školovanje započela sam u osnovnoj školi "Milan Lang" u Bregani te ga nastavila 2005. godine u općoj gimnaziji "Antun Gustav Matoš" u Samoboru. Nakon završenog srednjoškolskog obrazovanja, 2009. upisujem preddiplomski studij matematike inženjerskog smjera na Prirodoslovno - matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija 2012. godine stječem akademski naziv sveučilišne prvostupnice te iste godine upisujem diplomski studij statistike na Prirodoslovno - matematičkom fakultetu.